Alignement textométrique des unités lexicales à correspondances multiples dans les corpus parallèles

Maria Zimina

EA 2290 SYLED, Université de la Sorbonne nouvelle – Paris 3 – 19, rue des Bernardins – 75005 Paris – France – zimina@msh-paris.fr

Abstract

The approach suggested in this article enables statistic identification of multiple word correspondences in bilingual texts aligned on phrase level. Corresponding lexical units are discovered through characteristic element computation in parallel contexts. The calculation undergoes systematic reiteration within a corpus. The exploratory results show that the use of quantitative methods in combination with bilingual text topography offers new means for automatic description of lexical equivalences.

Résumé

Dans les corpus bilingues alignés au niveau de la phrase, le repérage de correspondances lexicales multiples peut être effectué sur des bases quantitatives. Fondée sur la pratique du calcul des spécificités, notre méthode explore parallèlement les contextes équivalents pour repérer des correspondances dans les emplois caractéristiques des différents types d'unités textuelles. La réitération systématique de ce processus dans le corpus, appuyée sur l'utilisation des principes de topographie bi-textuelle, offre de nouveaux moyens automatisés pour l'extraction et la description d'équivalences lexicales.

Mots-clés : corpus parallèles, alignement lexical, correspondances de traduction, textométrie, topographie bi-textuelle.

1. Les corpus parallèles alignés

Dans le contexte multilingue, les corpus de documents parallèles sont constitués de plusieurs volets qui correspondent chacun à une version d'un même texte dans plusieurs langues différentes. Il s'agit, en général, de textes sources et de leurs traductions (effectuées par des traducteurs humains) présentés sous forme électronique ou de textes dont chacun est une traduction de l'autre sans qu'il soit possible de déterminer lequel a servi de source. Le traitement de corpus parallèles suppose une phase préalable d'*alignement*, c'est-à-dire la mise en correspondance d'unités textuelles de différents types entre chacun des volets du *bi-texte*.

De nombreux travaux ont montré l'utilité des corpus parallèles alignés pour le développement des applications de traitement automatique des langues (Isabelle et Warwick-Armstrong, 1993 ; Véronis, 2000). La création automatique des lexiques bilingues alignés permet de normaliser la terminologie dans un champ de communication. Elle est indispensable à la maintenance de la documentation technique multilingue. Les travaux sur les corpus alignés contribuent aussi au développement des moteurs de recherche multilingues sur le Web. Pour le traducteur, les corpus alignés fournissent un accès au savoir-faire de la communauté lorsqu'il s'agit de traduire une expression pour laquelle les solutions proposées par des ouvrages de référence (dictionnaires, bases de données terminologiques, etc.) sont insatisfaisantes. En lexicographie, l'alignement des données de traduction met en lumière la spécificité et la

richesse du vocabulaire bilingue. Les corpus alignés permettent de découvrir des usages et des expressions ne figurant pas encore dans les dictionnaires.

2. Alignement lexical : problèmes et enjeux

Les comptes-rendus d'expériences publiés récemment décrivent des algorithmes permettant d'apparier les phrases d'un corpus parallèle avec un taux de réussite élevé (Véronis, 2000)[1]. En revanche, l'*alignement lexical*, c'est-à-dire la mise en correspondance de mots et locutions entre les deux volets d'un corpus parallèle demeure un problème difficile. Lors de l'alignement de ce type d'unités, il faut tenir compte de plusieurs phénomènes complexes liés, notamment, à la détection des emplois polysémiques de mots et de leur fonctionnement dans des séquences figées et locutions dont la traduction varie selon le contexte.

3. Corpus *Convention* : navigation textométrique

3.1. *Unités lexicales à correspondances multiples*

Lorsqu'il s'agit de mots dotés d'un large éventail de sens, les correspondances traductionnelles entre les deux volets d'un corpus parallèle forment un réseau complexe. Les unités lexicales d'un volet peuvent recevoir plusieurs traductions dans l'autre volet. Pour illustrer ce type d'équivalences, nous emprunterons des exemples à un corpus de textes juridiques anglais-français de la *Convention de sauvegarde des droits de l'homme et des libertés fondamentales*, désormais ***Convention*** (cf. *Tableau 1*)[2].

volet français	volet anglais
conv_a0_p1-1 1 Les gouvernements signataires , membres du Conseil de l'Europe, **conv_a0_p2-1 3** Considérant la Déclaration universelle des Droits de l'Homme, proclamée par l'Assemblée générale des Nations Unies le 10 décembre 1948 ; /…/ **conv_a0_p7-1 13** Sont convenus de ce qui suit /…/	**conv_a0_p1-1e 2** The governments signatory hereto, being members of the Council of Europe, **conv_a0_p2-1e 4** Considering the Universal Declaration of Human Rights proclaimed by the General Assembly of the United Nations on 10th December 1948; /…/ **conv_a0_p7-1e 14** Have agreed as follows /…/
Guide de lecture : Chaque couple de phrases équivalentes est introduit par un code. Les numéros indiquent (dans l'ordre) : le type de document (convention, protocole, etc.) ; le numéro de l'arrêt ; la partie de l'arrêt ; le numéro de la section et/ou du paragraphe ; le numéro de la phrase dans le corpus, précédé par la lettre « e » pour les phrases en anglais.	

Tableau 1. Corpus Convention *(extrait)*

Dans le corpus ***Convention***, le mot français *mort* (F=44) est le plus souvent traduit en anglais par *death* (F=26). On rencontre cette correspondance au sein des équivalences lexicales *la*

[1] Les systèmes actuels d'alignement des phrases de textes parallèles multilingues ont fait récemment l'objet d'une étude d'évaluation menée au sein du projet ARCADE. Les résultats de l'étude témoignent d'avancées méthodologiques importantes dans les techniques d'alignement des phrases. Lorsque les textes ne présentent pas de divergences importantes au niveau structurel (pas d'omissions, etc.), le taux de précision des systèmes évalués est estimé, en moyenne, à 98,5 % (cf. Véronis, 2000).

[2] Le corpus a été constitué à partir des documents contenus dans la Convention, d'une douzaine de protocoles, et d'une série d'arrêts rendus par la Cour européenne des Droits de l'Homme de Strasbourg en 1995. Chaque volet du corpus compte approximativement 300 000 mots graphiques. On peut trouver les textes de la Convention en anglais et en français sur le site officiel de la Cour Européenne des Droits de l'Homme : http://www.echr.coe.int.

peine de mort – death penalty, *les circonstances de la mort – the circumstances of the death*, *un danger de mort – a risk of death* et beaucoup d'autres. Cependant, l'équivalence *mort – death* n'est pas préservée dans les contextes suivants :

volet français	volet anglais
La mort n'est pas considérée comme infligée en violation de cet article /…/.	**Deprivation of life** shall not be regarded as inflicted in contravention of this article /…/
/…/ l'article vise certes les cas où **la mort** a été infligée intentionnellement, mais ce n'est pas son unique objet.	/…/ this provision extends to, but is not concerned exclusively with, intentional **killing**.

(Corpus ***Convention***)

Dans les deux couples de phrases ci-dessus le mot *mort* est utilisé dans le sens « fin provoquée de la vie », répertorié, par exemple, par le dictionnaire *Le Robert* (édition 1996). L'absence de cette signification dans l'univers sémantique du mot anglais *death* explique la présence des équivalences *mort – deprivation of life*, *mort – killing* dans le corpus. Comme on va le voir ci-dessous, une description automatique de correspondances traductionnelles multiples peut être envisagée si l'on fait appel à des méthodes d'analyse quantitative.

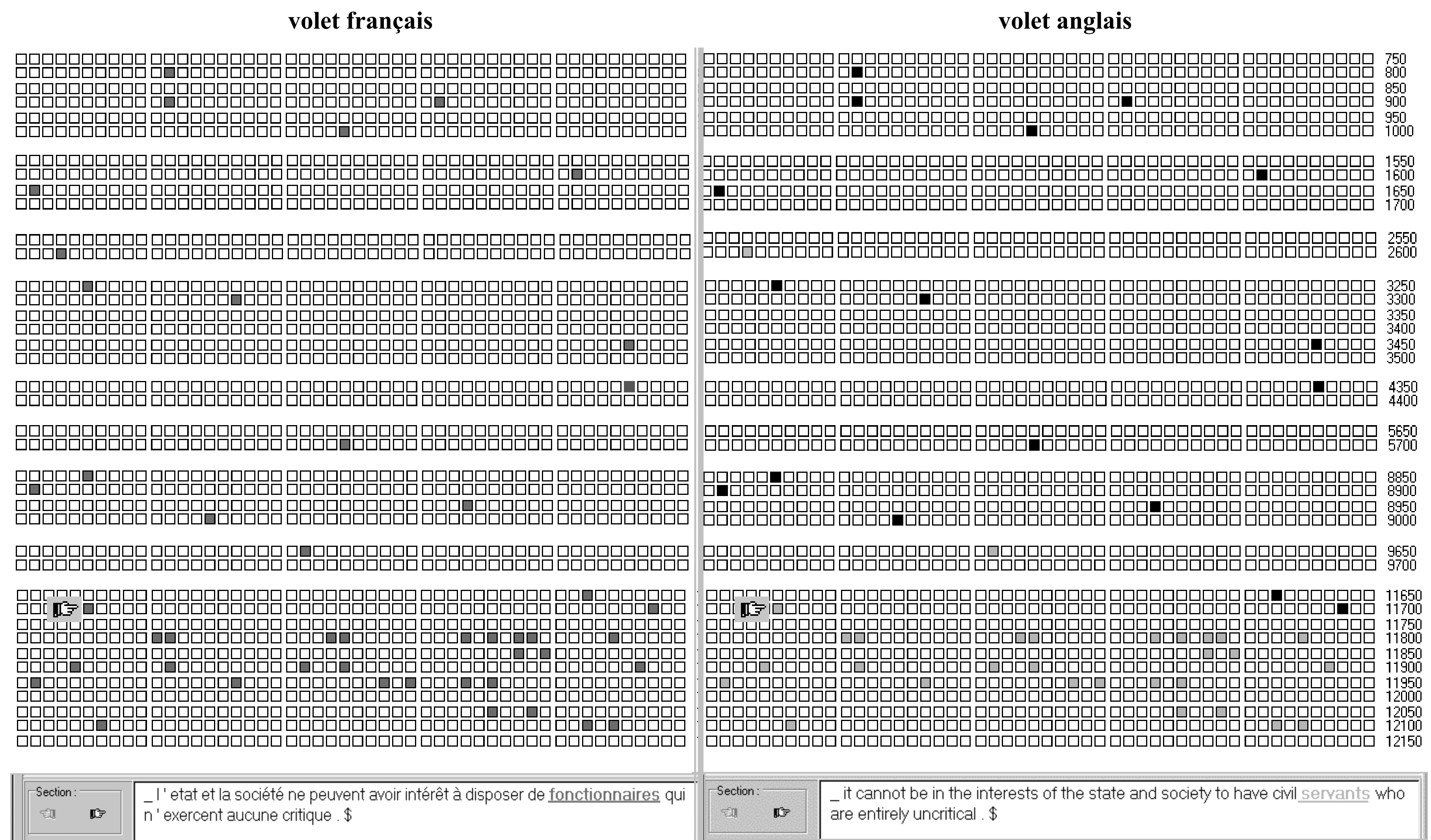
3.2. *Topographie bi-textuelle et outils statistiques d'appariement*

Les travaux récents montrent qu'il est possible de faire appel à des principes d'*analyse textométrique*[3] et, notamment, à une *représentation topographique*[4] du texte pour l'appariement des mots et des syntagmes des corpus parallèles (cf. Lamalle et Salem, 2002 ; Zimina, 2000 et 2002). La cartographie des *présence-absence* de correspondances bilingues au sein des traductions fournit des moyens automatisés pour le recensement des équivalences. Dans ce qui suit, nous allons décrire une méthode d'analyse permettant la découverte de correspondances lexicales entre les deux volets parallèles pour des mots qui possèdent plusieurs traductions au sein d'un même corpus. Une approche hybride qui allie la *topographie textuelle* et l'*analyse des spécificités*[5] est à l'origine de cette méthode. La recherche de correspondances lexicales s'appuie sur l'alignement du corpus au niveau de la phrase. L'exploration débute par le marquage au fil du texte dans l'un des volets du corpus bilingue d'un sous-ensemble d'occurrences correspondant à un *type* quelconque (*forme graphique*, *lemme*, *segment répété*, etc.). Le repérage des phrases correspondantes dans l'autre volet permet de construire deux fragments de texte équivalents qui seront confrontées dans chacun des cas au reste du corpus à des fins de comparaison.

[3] Sur la description de pratiques de la textométrie, on consultera Heiden (2002).

[4] La *topographie textuelle* a pour objectif une localisation graphique des phénomènes mis en évidence par l'étude statistique.

[5] Le repérage des *spécificités* ou vocabulaires caractéristiques met en évidence, pour un groupe de phrases donné, les unités dont la fréquence connaît une variation importante dans ce fragment de texte. Fondée sur le *model hypergéométrique*, la méthode des spécificités permet d'effectuer une comparaison entre l'ensemble du corpus (T) et l'échantillon des contextes contenant l'unité pôle (t). En fonction de la fréquence globale des unités attestées dans ce fragment (F) et de leur fréquence locale (f), on leur affecte un indice de spécificité. Le diagnostic est fourni sous la forme *±Exx* où le signe indique un *sur-emploi* ou un *sous-emploi* de l'unité et la valeur indique son degré de spécificité, cf. (Lafon, 1984 ; Lebart et Salem, 1994).



Lexico3

Figure 2 : Un extrait de la carte des phrases *appariées issues du corpus* Convention

Guide de lecture de la figure 2 : Les fonctionnalités développées au sein du logiciel ***Lexico3*** permettent d'envisager la visualisation de l'alignement des phrases d'un corpus parallèle à l'aide d'une *carte des sections*. Sur la figure 2, l'appariement des phrases d'un extrait du corpus ***Convention*** est matérialisé par des carrés positionnés sur deux colonnes. La ventilation de la forme *fonctionnaires* est représentée par les carrés gris foncé ■. Dans le fragment équivalent en anglais, les carrés gris clair ■ correspondent aux occurrences de l'une de ces traductions en anglais, la forme *servants*. La sélection des phrases activée sur la carte est matérialisée par des carrés noirs ■. Elle correspond à des phrases du volet anglais où la forme *fonctionnaires* reçoit d'autres traductions que *servants*.

Note : La description des fonctionnalités de la *carte des sections* prend en compte l'état final d'un ensemble de procédures en cours de développement qui devraient être disponibles dans la prochaine version de ***Lexico3***.

3.2.1. *Construction de la carte des sections*

Les fonctionnalités développées au sein du logiciel ***Lexico3*** (Lamalle *et al.*, 2003) permettent à l'utilisateur de visualiser une *carte des sections*[6] (ex.: phrases ou paragraphes du corpus), puis de constituer une sélection arbitraire de sections dont on étudiera ensuite le vocabulaire spécifique. L'utilisateur dispose d'un ensemble d'outils permettant de choisir (à partir du dictionnaire, du *garde-mots*[7], de la liste des *segments répétés*, etc.) un *type* sur lequel portera son exploration. Après avoir sélectionné le type, il est possible de le faire glisser sur la carte (*glisser/déposer*). La ventilation du type étudié devient alors visible. Les sections dans lesquelles il est présent sont coloriées[8]. Ce processus peut être réitéré.

Pour illustrer notre propos, nous considérerons les phrases du corpus ***Convention*** qui contiennent la forme *fonctionnaires* (F=49) et le sous-ensemble de phrases équivalentes en anglais. Sur la figure 2, l'appariement des phrases du corpus est matérialisé par des carrés positionnés sur deux colonnes. Les carrés sont liés et toute sélection dans un volet est automatiquement répercutée dans l'autre. La ventilation de la forme française *fonctionnaires* est représentée par des carrés gris foncé. L'appariement des phrases étant représenté sur la carte, il est possible d'envisager l'activation de la sélection parallèle des phrases équivalentes dans le volet anglais[9].

3.2.2. *Les outils textométriques de la carte des sections*

Le calcul des *spécificités* permet de sélectionner parallèlement pour chacun des fragments

[6] La ***carte des sections*** permet une visualisation du corpus découpé en sections par la promotion d'un (ou de plusieurs) caractères particuliers au statut de délimiteurs de section (Lamalle *et al.*, 2003). Dans le cas des corpus parallèles, le découpage en sections peut être effectué parallèlement, en s'appuyant sur des codes attribués aux phrases en correspondance.

[7] Le *garde-mots* est une fonctionnalité de ***Lexico3*** permettant de mémoriser formes, segments, etc. pour une utilisation ultérieure. Pour stocker un type dans le *garde-mots* il suffit de le faire glisser sur l'icône de cette fonctionnalité.

[8] Il est possible d'obtenir un coloriage plus ou moins sombre des sections en cochant d'abord la case *seuil*. Cette fonctionnalité permet de régler deux seuils en probabilités qui entraîneront l'intensité du coloriage sur la carte.

[9] La réalisation informatique de la sélection parallèle des sections correspondantes est actuellement en cours de développement. Elle s'appuie sur les fonctionnalités existantes de la *carte des sections* de ***Lexico3*** et utilise des principes formels de la segmentation parallèle des corpus bilingues où chaque couple de phrases équivalentes est introduit par le même code.

ainsi constitués une série d'unités textuelles caractéristiques de ces fragments[10]. Les listes des spécificités s'affichent dans les fenêtres des deux côtés de la *carte des sections bi-textuelle*. Le *Tableau 3* présente quelques-unes de ces unités spécifiques mises en évidence par cette méthode. On constate sur ce tableau que les unités bilingues issues de l'exploration constituent bien des correspondances de traduction. Ainsi, la forme française *fonctionnaires* (*spec.*+E109), ayant servi de point d'entrée pour la construction de l'échantillon de phrases pour cette analyse, peut être appariée avec la forme *servants* (*spec.*+E55) et le segment répété *civil servants* (*spec.*+E52) qui se sont révélés les plus caractéristiques du fragment anglais (cf. *Tableau 3*).

forme / segment	**Frq. Tot.**	**frq. locale**	**Spec.**	**forme / segment**	**Frq. Tot.**	**frq. locale**	**Spec.**
fonctionnaires	49	49	+E109	servants	50	31	+E55
les fonctionnaires	14	14	+E31	civil servants	46	29	+E52
des fonctionnaires	14	14	+E31	civil	304	41	+E40
de loyauté	36	14	+E22	loyalty	43	14	+E20
loyauté	42	14	+E21	duty	109	15	+E16
de loyauté politique	22	10	+E17	political loyalty	25	10	+E16
loyauté politique	24	10	+E16	of political	29	10	+E15
de fonctionnaires	7	7	+E16	duty of	45	11	+E15
obligation de loyauté politique	15	8	+E14	officers	38	10	+E14
obligation de loyauté	21	8	+E13	duty of political loyalty	23	9	+E14
				service	110	14	+E14
				civil service	58	11	+E13
				CONSTITUTIONAL	146	14	+E13

Guide de lecture : Le tableau représente un extrait de la liste des spécificités positives majeures calculées pour l'échantillon des phrases où est attestée la forme *fonctionnaires* et le fragment correspondant en anglais.

Tableau 3. Spécificités majeures pour la sélection bi-textuelle

La fréquence globale de la forme *fonctionnaires* (F=49) est supérieure à celle de la forme anglaise *servants* dans le fragment (f_{locale}=31). Nous pouvons en conclure que la forme *fonctionnaires* reçoit d'autres traductions dans le corpus. Pour découvrir l'ensemble des équivalences lexicales correspondant à la forme-pôle, on soumet au calcul des spécificités les seules phrases du fragment anglais dans lesquelles la forme *servants* est absente (cf. *Figure 2*). La réitération du calcul des spécificités dans ce sous-ensemble de phrases met en évidence une série d'unités les plus caractéristiques de ce nouveau fragment :

[10] Actuellement, les outils statistiques de la *carte des sections* rendent possible une sélection automatique des sections dans lesquelles le type étudié est présent (c'est cet ensemble de sections que l'on compare à l'ensemble du corpus).

forme / segment	Frq. Tot.	frq. locale	Spec.
officers	38	10	+E19
officials	16	7	+E16
senior	18	6	+E13
senior police	5	4	+E11
police	216	9	+E10
senior police officers	3	3	+E09

Le retour au contexte confirme que les unités *officers* (*spec.*+E19), *senior police officers* (*spec.*+E09), *officials* (*spec.*+E16), constituent bien des traductions de la forme *fonctionnaires* (au même titre que la forme *servants* et le segment *civil servants* découverts précédemment), (cf. *Tableau 4*). L'exploration contextuelle s'appuie sur les outils de navigation textométrique de la *carte des sections*. Sur la figure 2, les boutons situés à gauche de la fenêtre de visualisation de la sélection (en forme de mains) permettent de passer, respectivement, à la section suivante/précédente ou l'occurrence suivante/précédente du type étudié. Pour explorer parallèlement les deux volets bilingues du corpus, les sections en correspondance sont liées. Toute sélection dans une fenêtre est répercutée dans l'autre.

volet français	volet anglais
/…/ l'introduction de procédures disciplinaires à l'encontre de **fonctionnaires**, en raison de leur engagement politique /…/, violerait la convention de l'organisation internationale du travail (oit) /…/	/…/ the institution of disciplinary proceedings against *civil* **servants** on account of their political activities /…/ breached international labour organisation (ilo) convention /…/
il s'agissait en fait d'un document destiné aux agents du bvd (*binnenlandse veiligheidsdienst*) et d'autres **fonctionnaires** appelés à accomplir des missions pour lui.	it was in fact a document intended for bvd (*binnenlandse veiligheidsdienst*) staff and other **officials** who carried out work for the bvd**.**
il dénonce les propos tenus lors de la conférence de presse par le ministre de l'intérieur et les hauts **fonctionnaires** de police qui l'accompagnaient.	he complained of the remarks made by the minister of the interior and the *senior police* **officers** accompanying him at the press conference.

Tableau 4. Retours au contexte (extrait)

3.3 *Analyse des résultats*

L'exploration du corpus ***Convention*** à l'aide de la topographie bi-textuelle a permis de découvrir les principales traductions de la forme-pôle *fonctionnaires* (F=49) : *officers* (f_{locale}=10), *officials* (f_{locale}=7) et *servants* (f_{locale}=31). Un léger écart entre la fréquence globale de cette forme-pôle et le cumul des fréquences locales de ces correspondances en anglais montre qu'il existe au moins un contexte pour lequel la traduction n'a pas été identifiée par notre exploration. Nous pouvons affiner nos constats à travers un retour au texte. Il suffit d'écarter toutes les phrases du fragment anglais où la forme *fonctionnaires* est traduite par *officers*, *officials* ou *servants*. Pour ce faire, on procède à la sélection des phrases du fragment anglais dans lesquelles ces trois formes sont absentes ou contenues en nombre inférieur au total d'occurrences de la forme *fonctionnaires* dans la phrase correspondante en français. Cette recherche aboutit à la localisation sur la carte du corpus du couple de phrases suivantes :

volet français	volet anglais
Aux termes de /…/ la loi-cadre sur les **fonctionnaires** des *länder* /…/ seul peut être nommé fonctionnaire celui qui « offre la garantie qu'il prendra constamment fait et cause pour le régime fondamental libéral et démocratique au sens de la loi fondamentale. »	by virtue of /…/ the **civil service** (general principles) act for the *länder*, appointments to the civil service are subject to the requirement that the persons concerned "satisfy the authorities that they will at all times uphold the free democratic constitutional system within the meaning of the basic law".

(Corpus ***Convention***)

Dans ces dernières phrases, la forme *fonctionnaires* a été traduite par *civil service*. Cette équivalence relève de la notion d'*équivalence contextuelle*. Sur le plan sémantique, il s'agit d'unités traductionnelles singulières qui nécessitent un traitement particulier. Il appartient à l'expert humain de s'appuyer sur les blocs alignés pour examiner dans le détail les parallèles et les divergences entre ce type de séquences : *la loi-cadre sur les fonctionnaires ~ the civil service (general principles) act*.

Conclusions

Au terme de cette étude, nous avons défini une approche qui permet d'accéder à la description automatique de relations de correspondance entre des unités polysémiques qui possèdent plusieurs traductions au sein du corpus bi-textuel. Privilégiant le point de vue textométrique, notre approche repose entièrement sur les ressources construites à base de corpus. Appuyée sur l'utilisation de la représentation topographique de textes alignés au niveau de la phrase, cette approche offre à l'utilisateur de nouveaux moyens informatisés pour explorer la structure des équivalences qui se forment au niveau des mots et des syntagmes dans les textes originaux et leurs traductions.

Références

Heiden S. (2002). *Weblex : Manuel Utilisateur. Version 4.1* (intermédiaire). UMR 8503, ENS Lettres et Sciences humaines : http://lexico.ens-lsh.fr/doc/weblex.pdf.

Isabelle P. et Warwick-Armstrong S. (1993). Les corpus bilingues : une nouvelle ressource pour le traducteur. In Bouillon P. et Clas A. (Eds), *La Traductique : Études et Recherches de traduction par ordinateur.* Les Presses de l'Université de Montréal : 288-306.

Lafon P. (1984). *Dépouillements et statistiques en lexicométrie.* Slatkine-Champion.

Lamalle C. et Salem A. (2002). Types généralisés et topographie textuelle dans l'analyse quantitative des corpus textuels. In *Actes des JADT 2002* : 403-412.

Lamalle C., Martinez W., Fleury S., Salem A., Fracchiolla B., Kuncova A. et Maisondieu A. (2003). *Lexico3 – Outils de statistique textuelle. Manuel d'utilisation*. Syled-CLA^{2}T, Université de la Sorbonne nouvelle – Paris 3 : http://www.cavi.univ-paris3.fr/Ilpga/ilpga/tal/lexicoWWW.

Lebart L. et Salem A. (1994). *Statistique Textuelle.* Dunod.

Véronis J. (Ed.) (2000). *Parallel Text Processing: Alignment and use of translation corpora.* Kluwer Academic Publishers.

Zimina M. (2000). Alignement de textes bilingues par classification ascendante hiérarchique. In *Actes des JADT 2000* : 171-178.

Zimina M. (2002). Repérages lexicométriques des équivalences à basse fréquence dans les corpus bilingues. In Véronis J. (Ed.), *Lexicometrica*, n° spécial Corpus alignés : http://www.cavi.univ-paris3.fr/lexicometrica/thema/thema6.htm.